An ontology-based disambiguation of terms

Sophia Ananiadou[1,2], Larisa N. Soldatova[3], BalaKrishna Kolluru[1,2]

[1]The University of Manchester, UK, [2]National Centre for Text Mining, UK, [3]Aberystwyth University, UK

One of the traditional applications of ontologies in text mining is the use of hierarchy for disambiguation of terms. For example, it may be not clear from the text if the term 'jaguar' refers to a car or an animal. If this term is defined in an ontology which is used for the disambiguation of the text, then a parent class for the class 'jaguar' should be able to provide an answer. However, a problem is that ontologies are not consistent with each other, and perhaps they never will be. Even OBO ontologies, which are designed to be orthogonal, are actually not orthogonal [1]. One way of solving the problem is to use an initial set of 'trusted' ontologies which are consistent with each other, and a set of rules to point other relevant ontologies. Such an approach can also solve a problem of selection appropriate ontologies to support annotation. Currently, BioPortal contains ~200 ontologies and the number is rapidly growing.

Within the JISC (Joint Information Systems Committee) funded project CheTA (Chemistry Using Text Annotations), the following consistent set of ontologies have been used for annotation of Chemistry papers: ChEBI, FIX, REX. However, depending on the goals of annotation i.e. identification of most popular methods for prediction of biological activity of compounds and extraction information about such methods, or i.e. reasoning about molecular descriptors and chemical diversity, it may be desirable to use specific domain ontologies. An ontology for drug discovery investigations (DDI) provides links to a number of relevant ontologies which can be used for specific goals [2]. For example, the DDI class 'QSAR' which is defined as a planned process provides two links: via the relations 'has specified input/output' to BODO (Blue Obelisk Descriptor Ontology) [3] and QSAR-ML [3] which specify the descriptors of compounds for QSAR methods. While annotating papers from the ART Corpus [4], the term 'QSAR' has been identified in the paper b410053k and the BODO and QSAR-ML have been invoked. This allowed to detect that another term from the text 'EVA' (EigenVAlue) is a synonym for the QSAR-ML descriptor BCUT ("Eigenvalue based descriptor noted for its utility in chemical diversity described by Pearlman et al.") which is defined as a molecular descriptor. Without the use of DDI and consequently QSAR-ML, the terms QSAR and EVA will be classified by OSCAR [5] as chemical compounds with a likelihood of 0,33 and 0.22 respectively.

Such an approach is also of value to ontology developers and for verification and updating of an ontology. In the considered example, the paper b410053k is more recent than the referenced in the QSAR-ML one. This suggests that the definition of the descriptor (which also can be extracted from the paper) has to be checked and perhaps updated. The detected synonym can be included into a Lexicon.

References:

[1] Ghazvinian,A., Noy,N.F. & Musen,M.A. (2010) How Orthogonal are the OBO Foundry Ontologies? *J. of BioMed Semantics* (in press).
[2] Qi et al. (2010) An Ontology of Description of Drug Design Investigations. *J. of Integrative Bioinformatics* **7(3)**:126.
[3] Spjuth et al (2010) Towards interoperable and reproducible QSAR analyses: Exchange of datasets. *J. of Chemoinformatics* **2:5.**
[4] ART Corpus. http://www.ukoln.ac.uk/projects/ART_Corpus/
[5] OSCAR. http://sourceforge.net/projects/oscar3-chem/